Evaluating, comparing and combining density forecasts using the KLIC with an application to the Bank of England and NIESR "fan" charts of inflation*

James Mitchell
National Institute of Economic and Social Research (NIESR)

Stephen G. Hall
Department of Economics, Leicester University and NIESR

September 12, 2005

Abstract

This paper proposes and analyses the Kullback-Leibler Information Criterion (KLIC) as a unified statistical tool to evaluate, compare and combine density forecasts. Use of the KLIC is particularly attractive, as well as operationally convenient, given its equivalence with the widely used Berkowitz Likelihood Ratio test for the evaluation of individual density forecasts that exploits the probability integral transforms. Parallels with the comparison and combination of point forecasts are made. This and related Monte-Carlo experiments help draw out properties of combined density forecasts. We illustrate the uses of the KLIC in an application to two widely used published density forecasts for UK inflation, namely the Bank of England and NIESR "fan" charts.

Keywords: Density forecasts; Forecast comparison; Forecast combination; Kullback-Leibler Information Criterion

*Address for correspondence: James Mitchell, National Institute of Economic and Social Research, 2 Dean Trench Street, Smith Square, London, SW1P 3HE, U.K. Tel: +44 (0) 207 654 1926. Fax: +44 (0) 207 654 1900. E-Mail: j.mitchell@niesr.ac.uk. We would like to thank two anonymous referees for helpful comments. Particular thanks to Kenneth Wallis for comments on an earlier related paper that have also helped us develop our ideas in this paper. Mitchell gratefully acknowledges financial support from the ESRC (Award Reference: RES-000-22-0610). All errors remain our own.

1 Introduction

Density forecasts, or more popularly "fan" charts, are being used increasingly in economics and finance since they provide commentators with a full impression of forecast uncertainty. They reflect the fact that point forecasts, namely the "central tendency" of the forecast, are better seen as the central points of ranges of uncertainty. Therefore it is not a question of a given point forecast proving to be right and another point forecast proving to be wrong. Users of forecasts may not be surprised if, for example, inflation turns out to be a little higher than the point forecast. Indeed they may not be very surprised if it turns out much larger.

More formally, density forecasts of inflation provide an estimate of the probability distribution of its possible future values. In contrast to interval forecasts, that state the probability that the outcome will fall within a stated interval such as inflation falling between 1% and 3%, density forecasts provide a complete description of the uncertainty associated with a forecast; they can be seen to provide information on all possible intervals. Density forecasts of inflation in the UK, for example, are now published each quarter both by the Bank of England in its "fan" chart and the National Institute of Economic and Social Research (NIESR) in its quarterly forecast, and have been for the last ten years. Density forecasts inform the user of the forecast about the risks involved in using the forecast for decision making. Indeed, interest may lie in the dispersion or tails of the density itself; for example inflation targets often focus the attention of monetary authorities to the probability of future inflation falling within some pre-defined target range while users of growth forecasts may be concerned about the probability of recession. Moreover, volatility forecasts, as measured by the variance, and other measures of risk and uncertainty, can be extracted from the density forecast.[1]

Accordingly there is a growing literature that has sought to evaluate density forecasts *ex post*; e.g. see Diebold et al. (1999), Clements & Smith (2000), Clements (2004) and Wallis (2004). The tool used to evaluate these density forecasts is based on the probability integral transform (pit) of the outturn with respect to the forecast densities. These pit's will be uniform (and for one-step ahead density forecasts also independently and identically distributed [IID]) when the forecast densities coincide with the densities of the data-generating-process [DGP]; see Diebold et al. (1998). Thus testing uniformity offers users a statistical method to evaluate density forecasts similar to how point forecasts are traditionally evaluated statistically *ex post* based on their root mean squared error (RMSE) relative to the subsequent outturn.

Despite the burgeoning interest in and evaluation of density forecasts in economics less attention has been paid to statistically both comparing and combining competing density forecasts. This stands in contrast to the well-developed literature for point forecasts. Statistical tests that explicitly compare the accuracy of one point forecast with another are firmly established for point forecasts. Diebold & Mariano (1995) [DM] tests and their

[1]For further discussion of the importance of providing measures of uncertainty surrounding a "central tendency" (the point forecast) see Granger & Pesaran (2000), Garratt et al. (2003) and for a review Tay & Wallis (2000).

various extensions, for example, are now used widely to test statistically whether two point forecasts are equally accurate assuming some, usually a quadratic, loss function. DM-type tests are then used routinely to select the "best" forecast from a potentially large set of competing forecasts. Additionally, it is well recognised both theoretically and empirically that combining competing individual point forecasts of the same event can deliver more accurate forecasts, in the sense of a lower root mean squared error (RMSE); see Bates & Granger (1969) and Stock & Watson (2004).

Therefore in an attempt to provide users of density forecasts with a comparable tool-kit to that routinely used to examine point forecasts we propose and analyse the Kullback-Leibler Information Criterion [KLIC] as a unified means of evaluating, comparing and combining competing density forecasts whether model-based or subjectively formed. The KLIC is a well-respected measure of 'distance' between two densities. It has been used widely for over fifty years in a number of related ways, although it has not been related to the evaluation of density forecasts and the pit's.[2]

We believe the unity offered by the KLIC as a tool to analyse density forecasts to be attractive. Although the pit has become the industry-standard, although not exclusive as we see below, statistical means of evaluating individual density forecasts, many different distributional tests have been used to test uniformity or *via* a transformation normality. Despite this apparent choice, these tests can all be related to the KLIC. In particular, following Bao et al. (2004), we consider how one of the popular tests, namely the Berkowitz (2001) Likelihood Ratio [LR] test, can be directly related to the KLIC. This facilitates not just evaluation of the density forecasts individually and their comparison, as discussed by Bao *et al.*, but also their combination. Since the true density is unknown, devising an equivalent LR evaluation test based on the pit's is computationally convenient. The KLIC can then be used to compare competing density forecasts; a test for equal predictive performance can be constructed. Based on a loss differential series, this test is a direct generalisation of tests of equal point forecast accuracy popularised by DM and extended by West (1996) and White (2000). It is also shown to be equivalent to devising a test of equal density forecast accuracy when the logarithmic scoring rule rather than the pit is used to evaluate the density forecasts; see Giacomini (2002).[3] These tests formalise previous attempts that have compared *via* visual inspection alternative density forecasts according to their relative distance to, say, the uniform distribution; e.g. see Clements & Smith (2000).

We then discuss how the KLIC offers a means of combining competing density fore-

[2]In particular, the KLIC is the basis for the Akaike model selection criterion (AIC). The AIC is employed frequently to rank alternative models according to how close they are to the true but unknown density that generated the data. Estimated from the maximised log-likelihood, the AIC offers an approximately unbiased estimate of the expected, relative KLIC distance between a given model and the true but unknown density, treated as a constant across competing models; for further discussion and references see Burnham & Anderson (2002), Chapters 2 and 6. Focus in this paper is on estimation of the KLIC using the pit's so that density forecast evaluation, comparison and combination is operational both with model-based and non model-based (subjective) density forecasts.

[3]Scoring rules examine the quality of density forecast by assigning a numerical score based on the forecast and the event or value that materialises.

casts, extending the insights of Bao *et al.* from density forecast comparison to combination. While Clements (2005) and Granger et al. (1989) have considered, respectively, the combination of event and quantile forecasts, that inevitably involve a loss of information compared with consideration of the 'whole' density, the combination of density forecasts has been neglected. Indeed Clements (2003) identifies this as "an area waiting investigation" (p.2). Recently, however, Hall & Mitchell (2004*b*) and Wallis (2005) have re-introduced the finite mixture distribution as a means of combining density forecasts.[4] Indeed the finite mixture distribution is a well understood and much exploited means of combining density forecasts. For example, the Survey of Professional Forecasters [SPF], previously the ASA-NBER survey, has essentially used it since 1968 to publish a combined density forecast of future GNP growth and inflation. Since respondents to the SPF supply density forecasts in the form of histograms the average or combined density forecast is defined as the mean density forecast across respondents.[5] Despite this long history, to-date little attention has been paid to how the weights on the competing density forecasts in the finite mixture should be determined. But as experience of combining point forecasts has taught us, irrespective of its performance in practice use of equal weights is only one of many options. For example, one popular alternative, the so-called regression approach, is to tune the weights to reflect the historical performance of the competing forecasts.

Density forecast combination with the weights determined by the KLIC is considered within the context of Bayesian Model Averaging (BMA). BMA offers a conceptually elegant means of conditioning on the entire set of density forecasts under consideration, rather than a single 'best' forecast. It accounts for uncertainty about what is the 'best' model.[6] In the BMA framework the combination weights are the model's posterior probabilities. The KLIC provides a natural means of estimating these weights since the best model according to the KLIC is the model with the highest posterior probability.

The plan of the remainder of this paper is as follows. We review the statistical evaluation of individual density forecasts in Section 2 and specifically the Berkowitz LR test in Section 2.1. In Section 3 we explain and discuss how the Berkowitz LR test can be re-interpreted as a test of whether the KLIC equals zero. Section 4 shows how the KLIC can be used to compare statistically the accuracy of two competing density forecasts. Relationships with related tests that have been proposed recently, that involve use of scoring rules or nonparametric estimation of the true density, are also considered. In Section 5

[4]Hall & Mitchell (2004*a*) offer an alternative approach of combining density forecasts. Following Morris (1974, 1977) and Winkler (1981) they adopt a Bayesian approach where competing densities are combined by a "decision maker" who views them as data that are used to update a prior distribution. Hall and Mitchell also distinguish between combining competing forecasts of various moments of the forecast density and directly combining the individual densities themselves, as with the finite mixture density.

[5]The SPF survey has been analysed by *inter alia* Zarnowitz & Lambros (1987), Diebold et al. (1999), Giordani & Söderlind (2003) and Clements (2005).

[6]Garratt et al. (2003) and Pesaran & Zaffaroni (2004) have also considered the combination of probability forecasts using BMA. In contrast to the approach proposed in this paper their weights rely on estimation of a statistical model. $KLIC$ weights determined by the pit's, as we see below, are operational not just with model-based but also subjective (e.g. survey-based) density forecasts. Moreover, based on pit's they have the attraction of being familiar to those used to evaluating density forecasts.

we propose and analyse the finite mixture density as a tool for the combination of density forecasts. Focus is on how KLIC weights can be used to choose the combination weights. We draw out in detail some properties of combining density forecasts in this manner. Parallels with the combination of point forecasts are made. This is important in beginning to understand the situations in which density forecast combination will deliver improved forecasts. We find that in contrast to the combination of point forecasts (with variance or RMSE minimising weights) density forecast combination may not help even in-sample.[7] Nevertheless, we try to offer advice to practitioners about the use of combined density forecasts. This is supplemented by two Monte-Carlo experiments designed to draw out the properties of the proposed method of density forecast combination. Section 6 considers an application to UK inflation. We illustrate the use of the KLIC as a tool to evaluate, compare and combine the Bank of England and National Institute of Economic and Social Research (NIESR) "fan" charts of inflation. These well-known density forecasts have been published in 'real-time' for over ten years. *Inter alia* this application lets us determine whether in practice improved density forecasts for inflation, one year ahead, might have been obtained if one had combined the Bank of England and NIESR "fan charts". Section 7 concludes.

2 Evaluation of density forecasts: a review

While there exist well established techniques for the *ex post* evaluation of point forecasts, often based around the RMSE of the forecast relative to the subsequent outturn, only recently has the *ex post* evaluation of density forecasts attracted much attention. Currently, following Diebold et al. (1998), the most widespread approach is to evaluate density forecasts statistically using the pit, itself a well-established result.[8] Diebold et al. (1998) popularised the idea of evaluating a sample of density forecasts based on the idea that a density forecast can be considered "optimal" if the model for the density is correctly specified. One can then evaluate forecasts without the need to specify a loss function. This is attractive as it is often hard to define an appropriate general (economic) loss function. Alternatively, we could focus on a particular region of the density, such as the probability of inflation being in its target range; see Clements (2004).

A sequence of estimated h-step ahead density forecasts, $\{g_{1t}(y_t)\}_{t=1}^{T}$, for the realisations of the process $\{y_t\}_{t=1}^{T}$, coincides with the true densities $\{f_t(y_t)\}_{t=1}^{T}$ when the sequence of

[7] Out-of-sample, of course, there is no guarantee even with point forecasts that combination using variance or RMSE minimising weights will help. Indeed many studies have found that equal weights perform better out-of-sample; e.g. see Hendry & Clements (2004) and Smith & Wallis (2005).

[8] This methodology seeks to obtain the most "accurate" density forecast, in a statistical sense. It can be contrasted with economic approaches to evaluation that evaluate forecasts in terms of their implied economic value, which derives from postulating a specific (economic) loss function; see Granger & Pesaran (2000) and Clements (2004).

pit's, z_{1t}, are uniform U(0,1) variates where:[9]

$$z_{1t} = \int_{-\infty}^{y_t} g_{1t}(u)du = G_{1t}(y_t);\ (t = 1, ..., T). \tag{3}$$

Furthermore when $h = 1$, $\{z_{1t}\}_{t=1}^{T}$ are both U(0,1) and IID. In other words, one-step ahead density forecasts are optimal and capture all aspects of the distribution of y_t only when the $\{z_{1t}\}$ are IID U(0,1). When $h > 1$ we should expect serial dependence in $\{z_{1t}\}_{t=1}^{T}$ even for correctly specified density forecasts. This is analogous to expecting dependence (a $MA(h-1)$ process) when evaluating a sequence of rolling optimal h-step ahead point forecasts or optimal fixed-event point forecasts; e.g. see Clements & Hendry (1998), pp. 56-62. There is not, however, a one-for-one relationship between the point forecast errors and z_{1t}.

By taking the inverse normal cumulative density function (c.d.f.) transformation of $\{z_{1t}\}$ to give, say, $\{z_{1t}^*\}$ the test for uniformity can be considered equivalent to one for normality on $\{z_{1t}^*\}$; see Berkowitz (2001). For Gaussian forecast densities with mean given by the point forecast, z_{1t}^* is simply the standardised forecast error (outturn minus point forecast divided by the standard error of the Gaussian density forecast). Testing normality is convenient as normality tests are widely seen to be more powerful than uniformity tests. However, testing is complicated by the fact that the impact of dependence on the tests for uniformity/normality is unknown, as is the impact of non-uniformity/normality on tests for dependence.

Consequently various single and joint tests of U(0,1)/N(0,1) and IID have been employed in empirical studies.[10] These include Kolmogorov-Smirnov, Anderson-Darling, Doornik-Hansen tests for U(0,1)/N(0,1), Ljung-Box tests and LM tests for IID, Hong, Thompson and Berkowitz Likelihood Ratio (LR) tests for both U(0,1)/N(0,1) and IID. For empirical examples see Clements & Smith (2000), Clements (2004) and Hall & Mitchell (2004*a*).

2.1 Berkowitz's LR test

For $h = 1$ Berkowitz (2001) proposes a three degrees of freedom LR test of the joint null hypothesis of a zero mean, unit variance and independence of z_{1t}^* against z_{1t}^* following a

[9] It is instructive to run through the proof here; see Diebold et al. (1998) for details. *Via* the 'change of variables formula', we know z_{1t} has the probability density function (p.d.f.):

$$h_t(z_{1t}) = \left|\frac{\partial G_{1t}^{-1}(z_{1t})}{\partial z_{1t}}\right| f_t(G_{1t}^{-1}(z_{1t})), \tag{1}$$

$$= \frac{f_t(G_{1t}^{-1}(z_{1t}))}{g_{1t}(G_{1t}^{-1}(z_{1t}))}, \tag{2}$$

where $g_{1t}(y_t) = \frac{\partial G_{1t}(y_t)}{\partial y_t}$ and $y_t = G_{1t}^{-1}(z_{1t})$. Therefore when $g_{1t}(.) = f_t(.)$, $h_t(z_{1t}) = 1$ for $z_{1t} \in (0,1)$ and z_{1t} is a U(0,1) variate $(t = 1, ..., T)$.

[10] Alternatively, graphical means of exploratory data analysis are often used to examine the quality of density forecasts; see Diebold et al. (1998) and Diebold et al. (1999).

first-order autoregressive $AR(1)$ process: $z^*_{1t} = \mu + \rho z^*_{1t-1} + \varepsilon_t$, where $Var(\varepsilon_t) = \sigma^2$. The test statistic LR_B is computed as

$$LR_B = -2\left[L(0,1,0) - L(\widehat{\mu}, \widehat{\sigma}^2, \widehat{\rho})\right], \tag{4}$$

where $L(\widehat{\mu}, \widehat{\sigma}^2, \widehat{\rho})$ is the value of the exact log-likelihood of a Gaussian $AR(1)$ model; e.g. see Hamilton (1994) [p.119; eq. 5.2.9].[11] Under the null $LR_B \sim \chi^2_3$.

A criticism of Berkowitz's LR test is the maintained assumption of normality; see Clements & Smith (2000) and Bao et al. (2004). It only has power to detect non-normality through the first two moments. Consequently some authors, such as Clements and Smith and Hall & Mitchell (2004*a*), have supplemented the Berkowitz test with a nonparametric normality test, such as the Doornik-Hansen test. But as Bao et al. (2004) explain one can still construct a Berkowitz type LR test without maintaining the normality assumption. In this paper we confine attention to the traditional Berkowitz test, with normality maintained. Nevertheless, the discussion below is also applicable if a more general test were used to evaluate the density forecasts. Indeed any test, based on a sample average, is appropriate; the KLIC can still be used as a tool to evaluate, compare and combine density forecasts. We leave it for future work to study these more general tests. All of the single and joint statistical tests employed on the pit's, referred to above, amount to a test for whether KLIC $= 0$.[12]

3 Relating Berkowitz's LR test to the KLIC

The test for equal predictive accuracy of two competing density forecasts can be related to the well-known Berkowitz (2001) LR test for the statistical adequacy of an individual density forecast. This involves, following the suggestion of Bao et al. (2004), re-interpreting the Berkowitz LR test as a test of whether the KLIC 'distance' between the true (unknown) density and the forecast density equals zero. We first detail the equivalence between the Berkowitz LR test reviewed in Section 2.1 and the KLIC test, before turning in Section 4 to the test of equal predictive accuracy of two density forecasts.

Let the time-series y_t be a realisation from the sequence of (unknown) DGPs $f_t(y_t)$ $(t = 1, ..., T)$.[13] Note that the process $f_t(y_t)$ $(t = 1, ..., T)$ may, or may not, deliver outturns $\{y_t\}_{t=1}^T$ that are covariance-stationary. Nevertheless, some restrictions will be required to ensure consistent estimation of the asymptotic variance of the mean of the loss differential d_t defined below.

The Kullback-Leibler information criterion 'distance' measure KLIC_{1t} between the true

[11] The test can be readily generalised to higher order AR models; squared (and higher power) lagged values of z^*_{1t} can also be included in the model in an attempt to pick up nonlinear dependence; see Berkowitz (2001).

[12] For example, in terms of the ensuing discussion, a test for the uniformity of $\{z_{1t}\}$ amounts to a test for whether KLIC $= 0$ since under uniformity $h_t(z_{1t}) = 1$; see (8) below.

[13] For notational convenience we do not distinguish between random variables and their realisations.

density $f_t(y_t)$ and a density forecast $g_{1t}(y_t)$ $(t = 1, ..., T)$ is defined as:[14]

$$\mathsf{KLIC}_{1t} = \int f_t(y_t) \ln \left\{ \frac{f_t(y_t)}{g_{1t}(y_t)} \right\} dy_t \text{ or} \tag{5}$$

$$\mathsf{KLIC}_{1t} = \mathsf{E}\left[\ln f_t(y_t) - \ln g_{1t}(y_t)\right]. \tag{6}$$

The smaller this distance the closer the density forecast is to the true density; $\mathsf{KLIC}_{1t} = 0$ if and only if $f_t(y_t) = g_{1t}(y_t)$. For a related discussion see Vuong (1989).

Assuming $f_t(y_t)$ is known, under some regularity conditions, $\mathsf{E}\left[\ln f_t(y_t) - \ln g_{1t}(y_t)\right]$ can be consistently estimated by KLIC_1, the average of the sample information on $f_t(y_t)$ and $g_{1t}(y_t)$ $(t = 1, ..., T)$:

$$\mathsf{KLIC}_1 = \frac{1}{T} \sum\nolimits_{t=1}^{T} \left[\ln f_t(y_t) - \ln g_{1t}(y_t)\right]. \tag{7}$$

But $f_t(y_t)$ is, even *ex post*, unknown.[15] As we discuss below some authors have tried to estimate it. Alternatively, following Bao et al. (2004), we invoke Berkowitz (2001) Proposition 2 and note the following equivalence:

$$\ln f_t(y_t) - \ln g_{1t}(y_t) = \ln p_t(z^*_{1t}) - \ln \phi(z^*_{1t}) = \ln h_t(z_{1t}), \tag{8}$$

where $z_{1t} = \int_{-\infty}^{y_t} g_{1t}(u)du$, $z^*_{1t} = \Phi^{-1} z_{1t}$, $p_t(.)$ is the unknown density of z^*_{1t}, $\phi(.)$ is the standard normal density and Φ is the c.d.f. of the standard normal.[16]

In other words, testing the departure of $\{z^*_{1t}\}_{t=1}^T$ from $\mathsf{IID\ N(0,1)}$ [or $\{z_{1t}\}_{t=1}^T$ from $\mathsf{IID\ U(0,1)}$] is equivalent to testing the distance of the forecasted density from the true (unknown) density $f_t(y_t)$. Following Bao et al. (2004), we believe that testing whether $p_t(.)$ is $\mathsf{IID\ N(0,1)}$ is both more convenient and more sensible than testing the distance between $g_{1t}(y_t)$ and $f_t(y_t)$ since we do not know $f_t(y_t)$. If we did then, trivially, forecasting would be easy. Furthermore estimation of $f_t(y_t)$ typically requires some restrictions to be explicitly placed on the heterogeneity of y_t (e.g. covariance stationarity). Although $p_t(.)$ is also unknown at least we know that when $g_{1t}(y_t)$ is correctly specified $p_t(.)$ is $\mathsf{IID\ N(0,1)}$. We can therefore consider general forms for $p_t(.)$ that can accommodate non-normality

[14]The best model according to the KLIC is the model with the highest posterior probability; see Fernandez-Villaverde & Rubio-Ramirez (2004).

[15]Conceptually we may wish also to go as far as denying the existence of a "true" model. Rather than seeking to identify the true model, we may view the aim of economic modelling as seeking to approximate the truth adequately using a parsimonious model so that reliable inference can be made about the truth using this model.

[16]Again via the 'change of variables formula' we know that if $h_t(z_{1t})$ is the p.d.f of z_{1t} then the p.d.f of $z^*_{1t} = \Phi^{-1} z_{1t}$ is given by:

$$h_t(z^*_{1t}) = h_t(\Phi z^*_{1t}) \left| \frac{\partial z_{1t}}{\partial z^*_{1t}} \right| \tag{9}$$

where $\left|\frac{\partial z_{1t}}{\partial z^*_{1t}}\right|$ is the Jacobian of the transformation. Therefore $p_t(z^*_{1t}) = h_t(z_{1t})\phi(z^*_{1t}) = \left(\frac{f_t(y_t)}{g_{1t}(y_t)}\right)\phi(z^*_{1t})$. Taking logs we complete the proof.

and dependence but that include IID N(0,1) as a special case.[17] On the other hand, when we specify $g_{1t}(y_t)$ there is no certainty that it accommodates $f_t(y_t)$.

To test the null hypothesis that $f_t(y_t) = g_{1t}(y_t)$ we exploit the framework of West (1996) and White (2000).[18] Consider the loss differential $\{d_t\}$:

$$d_t = [\ln f_t(y_t) - \ln g_{1t}(y_t)] = [\ln p_t(z^*_{1t}) - \ln \phi(z^*_{1t})]\,;\ (t = 1, ..., T). \tag{10}$$

The null hypothesis of the density forecast being correctly specified is then

$$H_0 : \mathsf{E}(d_t) = 0 \Rightarrow \mathsf{KLIC}_1 = 0 \tag{11}$$

The sample mean $\overline{d}$ is defined as:

$$\overline{d} = \mathsf{KLIC}_1 = \frac{1}{T}\sum\nolimits_{t=1}^{T} [\ln p_t(z^*_{1t}) - \ln \phi(z^*_{1t})]\,. \tag{12}$$

To test a hypothesis about d_t we know $\overline{d}$ has *via* a central limit theorem, under appropriate assumptions, the limiting distribution

$$\sqrt{T}(\overline{d} - \mathsf{E}(d_t)) \xrightarrow{d} \mathsf{N}(0, \Omega), \tag{13}$$

where a general expression, allowing for parameter uncertainty, for the covariance matrix Ω is given in West (1996) Theorem 4.1. Under certain conditions parameter uncertainty is asymptotically irrelevant and Ω reduces to the long run covariance matrix associated with $\{d_t\}$ or 2π times the spectral density of $d_t - \mathsf{E}(d_t)$ at frequency zero.[19] This is the result seen in Diebold & Mariano (1995).[20] This long run covariance matrix S_d is defined as $S_d = \gamma_0 + 2\sum\nolimits_{j=1}^{\infty} \gamma_j$, where $\gamma_j = \mathsf{E}(d_t d_{t-j})$. *HAC* robust (to serial correlation and heteroscedasticity) estimators of the variance of the loss differential d_t can be used to estimate S_d consistently. See White (1984), Chapter 6, for discussion of the conditions necessary for consistent estimation.

[17] In theory $p_t(.)$ should be as general as possible to reflect the true density of z^*_{1t}. Bao et al. (2004) suggest that ε_t be allowed to follow a more general distribution than the Gaussian, which as seen in Section 2.1 was considered by Berkowitz. Specifically Bao et al. propose the use of a semi nonparametric density.

[18] West and White develop a general framework for inference about predictive ability that *inter alia* can account for parameter uncertainty, when relevant.

[19] There appear to be two distinct sources of parameter uncertainty. West (1996) provides conditions for asymptotic irrelevance pertaining to the first. This first source of uncertainty is when the density forecasts $g_{1t}(y_t)$ are generated conditional on a set of (estimated) parameters. The second source of parameter uncertainty arises from the fact that the loss differential series d_t may itself require parameters (e.g. μ, ρ and σ) in $p_t(z^*_{1t})$ to be estimated; see below. In general, the work of West (and co-authors) has shown that ignoring parameter uncertainty distorts statistical tests employed on forecasts and forecast errors. Therefore future work is required to investigate the effect of parameter uncertainty on the properties of the KLIC, especially in small samples.

[20] When $\{d_t\} = \left\{e^2_{1t} - e^2_{2t}\right\}$, where e_{it} is the point forecasting error for forecast i, the test reduces to a DM test of equal point forecast accuracy as measured by the RMSE.

Alternatively to this asymptotic test one could construct, along the lines of White's "bootstrap reality check", a (small-sample) test based on the bootstrap. This would involve re-sampling the test statistic $\overline{d} = \mathsf{KLIC}_1$ by creating R bootstrap samples from $\{d_t\}_{t=1}^{T}$ accounting for any dependence by using the so-called stationary bootstrap that resamples using blocks of random length.

The test statistic KLIC_1 is proportional (by a factor $2T$) to the LR test of Berkowitz (2001), assuming Gaussianity of ε_t; i.e. $LR_B = 2T\mathsf{KLIC}_1$. In terms of (12) Berkowitz's test, see (4), corresponds to assuming $p_t(z_{1t}^*) = \phi\left[\left(z_{1t}^* - \mu - \rho z_{1t-1}^*\right)/\sigma\right]/\sigma$, where ϕ is the (standardised) Gaussian density; i.e.

$$LR_B = 2\sum\nolimits_{t=1}^{T}\left[\ln p_t(z_{1t}^*) - \ln \phi(z_{1t}^*)\right]. \tag{14}$$

Therefore, asymptotic critical values from the chi-squared distribution can be used to test the null hypothesis, (11); these of course should be divided by $2T$. As stressed already this test only has power to detect non-normality through the first two moments. Of course, as Bao et al. suggest (see footnote 17), we could overcome this criticism by considering more general forms for $p_t(z_{1t}^*)$. As also explained by Bao et al. rather than evaluating the performance of the 'whole' density we can also evaluate in any regions of particular interest.

4 A test of equal predictive accuracy of two density forecasts using the KLIC

Extending the discussion above, a test for equal density forecast accuracy of two competing (non-nested) density forecasts $g_{1t}(y_t)$ and $g_{2t}(y_t)$, both of which may be misspecified, is then constructed based on $\{d_t\}_{t=1}^{T}$, where:[21]

$$d_t = \left[\ln f_t(y_t) - \ln g_{1t}(y_t)\right] - \left[\ln f_t(y_t) - \ln g_{2t}(y_t)\right], \tag{15}$$

$$d_t = \ln g_{2t}(y_t) - \ln g_{1t}(y_t), \tag{16}$$

$$d_t = \left[\ln p_t(z_{1t}^*) - \ln \phi(z_{1t}^*)\right] - \left[\ln p_t(z_{2t}^*) - \ln \phi(z_{2t}^*)\right]. \tag{17}$$

The null hypothesis of equal accuracy is then

$$H_0 : \mathsf{E}(d_t) = 0 \Rightarrow \mathsf{KLIC}_1 - \mathsf{KLIC}_2 = 0. \tag{18}$$

The sample mean $\overline{d}$ is defined as:

$$\overline{d} = \frac{1}{T}\sum\nolimits_{t=1}^{T}\left[\left[\ln p_t(z_{1t}^*) - \ln \phi(z_{1t}^*)\right] - \left[\ln p_t(z_{2t}^*) - \ln \phi(z_{2t}^*)\right]\right]. \tag{19}$$

[21] The testing approach developed here extends Vuong (1989) to time-series data and is appropriate out-of-sample. For IID (in-sample) data, Vuong proposed a statistical test for whether two models, say, are equally close to the true model, where distance is measured by KLIC. Traditionally this distance is proxied by the LR test statistic, perhaps corrected in-line with the parsimony of each model as reflected by the Akaike or Schwarz criterion. As in classical nested hypothesis testing the results of this test indicate the statistical evidence for a given model based on its goodness-of-fit.

Following (13) again a test can be constructed based on the fact that we know $\overline{d}$ has, under appropriate assumptions, the limiting distribution

$$\sqrt{T}(\overline{d} - \mathsf{E}(d_t)) \xrightarrow{d} \mathsf{N}(0, \Omega). \tag{20}$$

This again reduces to a DM type test in the absence of parameter uncertainty: $\overline{d}/\sqrt{\frac{S_d}{T}}$ $\xrightarrow{d}$ $\mathsf{N}(0,1)$. Additionally one could also exploit White's "bootstrap reality check" and construct an alternative to this asymptotic test based on the bootstrap. Also as suggested by White (2000) the test of equal predictive accuracy (18) can be readily extended to multiple (greater than two) models.

In fact to avoid having to postulate an unknown density $p_t(.)$ it is more convenient to couch the test in terms of (16) rather than (17). In this case we can clearly see that this test is equivalent to that proposed by Giacomini (2002). Giacomini proposes tests that can be used to compare the accuracy of density forecasts where evaluation is based on the logarithmic score, e.g. $\ln g_{1t}(y_t)$, rather than the pit's. Therefore the scoring rules approach of Giacomini is perhaps less restrictive than one might have originally thought; its focus on scoring rules might mistakenly lead one to conclude that it is divorced from the pit literature, but as seen here there is a strong relationship between the two. The use of a logarithmic scoring rule is then seen to be less arbitrary than again it might have initially appeared.

Related approaches to compare density forecasts statistically have been proposed by Sarno & Valente (2004) and Corradi & Swanson (2004). Rather than use the KLIC measure of 'distance' they rely on the integrated square difference between the forecast density and the true density (Sarno and Valente) and the mean square error between the c.d.f. of the density forecast and the true c.d.f., integrated out over different quantiles of the c.d.f. (Corradi and Swanson). But rather than rely on the pit's in both cases they estimate the true density or c.d.f. empirically. For example, Sarno and Valente implicitly assume $f_t(y_t) = f(y_t)$ $(t = 1, ..., T)$ and estimate $f(y_t)$ using the kernel estimator $\widehat{f}(y_t) = \frac{1}{Th}\sum_{i=1}^{T} K(\frac{y_i - y_t}{h})$, where $K(.)$ is the kernel function and h a smoothing parameter. Both these approaches for comparing density forecasts are related to Li & Tkacz (2001) who propose to evaluate an individual density forecast based on its (integrated squared) distance from a nonparametric estimate of the true density function. While one distance measure is not more natural than another, the KLIC can be readily computed for subjective (e.g. survey based) as well as model based density forecasts and, in our opinion, benefits from being based on the pit's and not relying on estimation of $f_t(y_t)$.

(18) is an *unconditional* test for equal forecast accuracy; see Giacomini & White (2004) [GW]. GW have developed more general *conditional* tests. These test which forecast will be more accurate at a future date rather than, as with the unconditional tests, testing which forecast was more accurate 'on average'. One could, for example, then recursively select at time t the best forecasting method for $t+1$. Conditional tests can be straightforwardly implemented in our framework. The null hypothesis of equal conditional forecast accuracy (for one-step ahead forecasts) amounts to testing $\mathsf{E}(d_t \mid h^*_{t-1}) = \mathsf{E}(h^*_{t-1} d_t) = 0$ $(t =$

$2, 3, ...$), where h_t^* is a vector of "test functions" which we set equal to $h_{t-1}^* = (1, d_{t-1})'$. The GW test statistic GW_T can be computed as the Wald statistic:

$$GW_T = T\left(T^{-1}\sum_{t=2}^{T} h_{t-1}^* d_t\right)' \widehat{\Sigma}_T^{-1}\left(T^{-1}\sum_{t=2}^{T} h_{t-1}^* d_t\right), \tag{21}$$

where $\widehat{\Sigma}_T$ is a consistent estimator for the asymptotic variance of $h_{t-1}^* d_t$ and $GW_T \xrightarrow{d} \chi_2^2$. GW note that a robust HAC estimator for this variance could be employed, as with DM-type tests, but they also explain that the sample variance is a consistent estimator when one exploits the fact that the null hypothesis implies $\left\{h_{t-1}^*, d_t\right\}_{t=2}^{T}$ is a martingale difference sequence. GW argue that this has the advantage of allowing the data $\{y_t\}$ to be heterogenous and characterised by arbitrary structural breaks at unknown points. Their test is also valid for nested models.

5 Combination of Density Forecasts using KLIC Weights

Rather than select a single 'best' forecast it can be felicitous to combine competing forecasts. This follows from appreciation of the fact that although one model may be 'better' than the others, we may not select it with probability one; we may not be sure that it is the best forecast. Therefore if we considered this single forecast alone we would be overstating its precision. We may better approximate the truth, and account for the uncertainty in model selection, by combining forecasts.

5.1 Bayesian Model Averaging

The Bayesian approach, so-called Bayesian Model Averaging (BMA), offers a conceptually elegant means of dealing with this model uncertainty. BMA is an application of Bayes' theorem; model uncertainty is incorporated into the theorem by treating the set of models S as an additional parameter and then integrating over S, where $S \equiv \{S_i, i = 1, ..., N\}$ and the models S_i are defined as continuous density functions $g_{it}(y_t)$ for the variable of interest y_t; for further discussion see Draper (1995).

The posterior density of the variable of interest y_t given 'data' Ω_t, $p_t(y_t \mid \Omega_t)$, is then defined as the weighted average of the predictive densities $g_{it}(y_t) = \Pr(y_t \mid S_{it}, \Omega_t)$, where the weights w_{it} are the model's posterior probabilities, $w_{it} = \Pr(S_{it} \mid \Omega_t)$:[22]

$$p_t(y_t \mid \Omega_t) = \sum_{i=1}^{N} w_{it} g_{it}(y_t);\ (t = 1, ..., T), \tag{22}$$

where $w_{it} \geq 0$ and $\sum_{i=1}^{N} w_{it} = 1$. $p_t(y_t \mid \Omega_t)$, or for expositional ease suppressing dependence on the 'data' Ω_t when defining the posterior probabilities equivalently $p_t(y_t)$, is the combined density forecast. Outside of the Bayesian paradigm, this finite mixture density

[22] All probabilities are implicitly conditional on the set of all models S under consideration.

is known as the "linear opinion pool". (22) satisfies certain properties such as the "unanimity" property (if all forecasters agree on a probability then the combined probability agrees also); for further discussion, and consideration of other properties see Genest & Zidek (1986) and Clemen & Winkler (1999). Wallis (2005) has also motivated $p_t(y_t)$ as the combined density forecast. Further descriptive properties of mixture distributions are summarised in Everitt & Hand (1981).

5.2 Characteristics of the combined density forecast

Inspection of (22) reveals that taking a weighted linear combination of the N individual densities can generate a combined density with characteristics quite distinct from those of the individual density forecasts. For example, if all the forecast densities are normal, but with different means and variances, then the combined density will be mixture normal. Mixture normal distributions can have heavier tails than normal distributions, and can therefore potentially accommodate skewness and kurtosis. If the true (population) density is non-normal we can begin to appreciate why combining individual density forecasts, that are normal, may mitigate misspecification of the individual densities. We explore further this issue in some simple Monte-Carlo experiments in Section 5.5.

Further characteristics of the combined density $p_t(y_t)$ can be drawn out by defining m_{it} and v_{it} as the mean and variance of forecast i's distribution at time t: $m_{it} = \int_{-\infty}^{\infty} y_t g_{it}(y_t) dy_t$ and $v_{it} = \int_{-\infty}^{\infty} (y_t - m_{it})^2 g_{it}(y_t) dy_t$; $(i = 1, ..., N)$. Then the mean and variance of (22) are given by:[23]

$$\mathsf{E}\left[p_t(y_t)\right] = m_t^* = \sum_{i=1}^{N} w_{it} m_{it}, \tag{23}$$

$$\mathsf{Var}\left[p_t(y_t)\right] = \sum_{i=1}^{N} w_{it} v_{it} + \sum_{i=1}^{N} w_{it} \left\{m_{it} - m_t^*\right\}^2. \tag{24}$$

(24) indicates that the variance of the combined distribution equals average individual uncertainty ("within" model variance) plus disagreement ("between" model variance).[24] But this result does not stand in contrast to that obtained when combining point forecasts where combination using "optimal" (variance or RMSE minimising) weights means the

[23] Related expressions decomposing the aggregate density (22), based on the 'law of conditional variances', are seen in Giordani & Söderlind (2003). This law states that for the random variables y_t and i: $V(y_t) = E[V(y_t|i)] + V[E(y_t|i)]$. For criticism see Wallis (2005).

[24] For further discussion of the relationship, if any, between dispersion/disagreement and individual uncertainty see Bomberger (1996).

RMSE of the combined forecast must be equal to or less than that of the smallest individual forecast; see Bates & Granger (1969) and for related discussion in a regression context Granger & Ramanathan (1984) [G-R]. This is because while density forecast combination increases the variance relative to its average across individuals, see (24), the variance or uncertainty of this variance (about the 'true' but unknown variance) need not rise and could fall. The weights used to combine will affect what happens. Similarly, while point forecast combination may or may not increase the mean forecast $\mathsf{E}\left[p_t(y_t)\right]$, its variance (about what we might consider the 'true' mean y_t) may or may not fall, again depending on the weights used. When optimal G-R weights are used we know that this variance will fall.

Indeed focusing on the entire density rather than a single moment, the combined density forecast does provide better predictive accuracy as measured by the logarithmic score when the weights used are the model's posterior probabilities; in turn this implies that the combined density forecast minimises the KLIC distance relative to $f_t(y_t)$. This follows from the fact that since $\mathsf{KLIC}_{it} \geq 0$, $\mathsf{E}(\ln p_t(y_t)) \geq \mathsf{E}(\ln g_{it}(y_t))$ $(i = 1, ..., N;$ $t = 1, ..., T)$; see Raftery et al. (1997). But crucially this does not imply that the predictive accuracy of any given moment improves.

5.3 Determination of the combination weights

The key practical problem we face when seeking to combine N density forecasts using (22) is how to determine w_{it}. Maintaining a Bayesian perspective w_{it} is given as:

$$w_{it} = \Pr(S_{it} \mid \Omega_t) = \frac{\Pr(\Omega_t \mid S_{it}) \Pr(S_{it})}{\sum_{i=1}^{N} \Pr(\Omega_t \mid S_{it}) \Pr(S_{it})}. \tag{25}$$

We assume uniform priors on $\Pr(S_{it})$ for all i $(i = 1, ..., N)$. Then define $B_t = \frac{w_{it}}{w_{jt}} = \frac{\Pr(\Omega_t|S_{it})}{\Pr(\Omega_t|S_{jt})}$ as the Bayes factor for model i against model j. B_t describes the contribution of the data towards the posterior odds: $\frac{\Pr(S_{it}|\Omega_t)}{\Pr(S_{jt}|\Omega_t)}$. The Bayes factor reflects how much the data will cause us to change our prior probabilities about each model. Note $B_t = \frac{\Pr(\Omega_t|S_{it})}{\Pr(\Omega_t|S_{jt})} = \frac{L(\Omega_t|S_{it})}{L(\Omega_t|S_{jt})}$, the relative likelihood L of model i versus model j. Therefore $B_t = \frac{g_{it}(y_t)}{g_{jt}(y_t)}$and $\ln B_t = \ln g_{it}(y_t) - \ln g_{jt}(y_t)$, the logarithmic score. Accordingly Good (1952) called the logarithmic Bayes factor the "weight of evidence".

At a practical level, following the discussion above, one can therefore move from KLIC_i to w_i. Note that $w_{it} = w_i$ since we rely on sample averages to estimate the KLIC; cf. (12). Of course on an out-of-sample basis these weights can be time-variant. Moving from KLIC_i to w_i is related to how one moves, say, from the Akaike criterion to Akaike

weights.[25] The KLIC weights w_i are defined as:

$$w_i = \frac{\exp(-\Delta_i)}{\sum_{i=1}^{N} \exp(-\Delta_i)} \ (i = 1, ..., N), \tag{26}$$

where $\Delta_i = \mathsf{KLIC}_i - \min(\mathsf{KLIC})$, where $\min(\mathsf{KLIC})$ is the minimum of the N different KLIC_i values, and $\sum_{i=1}^{N} w_i = 1$. Therefore $\Delta_i = 0$ for the best density and is positive for the other density forecasts; the larger Δ_i the less plausible is density i as the best density. From (26) we see that, consistent with our discussion about the log Bayes factor, $w_1/w_2 = \mathsf{E}\left[g_{1t}(y_t)/g_{2t}(y_t)\right]$, the expected ratio of the two density forecasts (in an in-sample context this ratio can be interpreted as the expected ratio of the likelihood functions of the models given the data: i.e. the relative strength of model 1 over model 2 since the likelihood of the i-th model $L_i \propto \exp(-\Delta_i)$, since $\ln f_t(y_t)$ is constant across models i). w_i can be interpreted as the probability that density forecast i is the most accurate density forecast in a KLIC sense.

Related work has used model selection criteria like the Akaike and Schwarz information criteria to proxy the posterior probabilities and define w_i; see Garratt et al. (2003) and Pesaran & Zaffaroni (2004).[26] These measure the relative statistical in-sample fit of the model. An advantage of KLIC weights is that since they do not rely on statistical estimation of a model as with Akaike and Schwarz weights but rather the pit's they are operational both with model and non-model (e.g. survey-based) density forecasts. Moreover, as we argue in this paper, we believe there is an elegance to a unified tool for the evaluation, comparison and combination of density forecasts that exploits the pit's.

Alternative approaches to the determination of w_i have been suggested. There is the data-driven approach of Hall & Mitchell (2004*b*) discussed further in Section 6.3.4. Granger & Jeon (2004) suggest a thick-modelling approach, based on trimming to eliminate the k% worst performing forecasts and then taking a simple average of the remaining forecasts. Most simply, equal weights, $w_i = 1/N$, have been advocated; e.g. see Hendry & Clements (2004) and Smith & Wallis (2005). Indeed they are used by the SPF when publishing their combined density forecasts. Markov Chain Monte-Carlo simulation methods have also been used; e.g. see Raftery et al. (1997).

5.4 Further characteristics of the combined density forecast

It is instructive to draw out some characteristics of these combination weights. For simplicity consider combining just two density forecasts so that $w_1 = \frac{\exp(-\Delta_1)}{\exp(-\Delta_1)+\exp(\Delta_2)} = \frac{\exp(-\mathsf{KLIC}_1)}{\exp(-\mathsf{KLIC}_1)+\exp(-\mathsf{KLIC}_2)}$. Then even if $\mathsf{KLIC}_1 \to 0$, $w_1 \nrightarrow 1$ unless $\mathsf{KLIC}_2 \to \infty$ also. This means that even if the true model is one of the two models under consideration, if the

[25] For a textbook exposition see Burnham & Anderson (2002), p. 75. These AIC weights can be interpreted from a Bayesian perspective as the probability that the model is the best approximation to the truth given the data; see Burnham & Anderson (2002), pp. 302-305.

[26] Minimising the Akaike criterion is approximately equivalent to minimising the expected Kullback-Leibler distance between the true density and the estimated density.

other model is 'close' to the truth, then a weight of one will not be given to the true model. This all follows from the uniform prior imposed on $\Pr(S_{it})$. A by-product of assuming a uniform prior and using the relative likelihoods (the data) to define the weights is that even if the true model is under consideration it will not receive a weight of one, unless the other models examined are very bad. To confront this problem one might consider some Bayesian learning type process whereby in the light of the empirical evidence for the model, as reflected by the KLIC, the prior is recursively (with the arrival of new data) updated. We leave this issue to future research. Alternatively, one might consider what we call, naïve BMA weights that although without obvious theoretical foundation do exhibit the property that $w_1 \to 1$ as $\mathsf{KLIC}_1 \to 0$, irrespective of the size of KLIC_2:

$$w_{1,naïve} = 1 - \frac{\mathsf{KLIC}_1}{\mathsf{KLIC}_1 + \mathsf{KLIC}_2}. \tag{27}$$

Encouragingly, these naïve weights ensure that the true model will receive a weight of unity.

It is also instructive to draw out some further characteristics of this density forecast combination method by relating it to the combination of point forecasts. Density forecast combination using (22) with weights (26) can, as explained, deliver 1,0 weights ($w_1 = 1$ and $w_i = 0$ for $\forall i \neq 1$) but this does not necessarily, in contrast to combination in the point forecast case, imply the forecast with a weight of 1 is "optimal" (in the sense of Diebold et al.; i.e. KLIC $= 0$). It just implies it is better than the other density forecast. Furthermore, when the weights are not 1,0 this does not imply, in contrast to the case of point forecasts combined following G-R, that combination will deliver improved density forecasts (in-sample). Only when the 'true' model is in the set of models under consideration (and the other models perform badly) will not only the weight on this model (the posterior probability of it being the "best" density forecast, in the sense of minimising the KLIC 'distance') be unity but the KLIC distance be zero.

The analogous case with point forecast combination appears to be testing for "conditional efficiency" (*encompassing*) of forecast 1 relative to forecast 2 (a zero coefficient on forecast 2 in the G-R regression) but not simultaneously "Mincer-Zarnowitz [M-Z] efficiency" (a unit coefficient on forecast 1 in the G-R regression).[27] With density forecasts the counterpart of M-Z efficiency is met only when the pit's associated with the combined (1,0) (i.e. the best individual) forecast are also IID U(0,1) (or the KLIC distance is zero). So to establish efficiency (in both senses) of the combined density forecast it is important to supplement examination of the weights (although we are not testing their significance statistically, accounting for their sampling variability) with a statistical test for IID U(0,1) or KLIC $= 0$. Density forecast combination *via* the linear opinion pool requires the weights to sum to unity. Future work is required to consider how we might simultaneously examine the density analogues of conditional and M-Z efficiency. Statistical testing of the significance of the KLIC weights, moving from combination to encompassing-type tests, also appears to remain an important area for research with density forecasts.

[27] Related discussion for forecast probabilities of events is provided by Clements (2005).

5.5 Monte-Carlo experiments

We consider two separate Monte-Carlo experiments to draw out further some properties of density forecast combination.

Let R denote the number of Monte-Carlo replications; we set $R = 500$. $T = 50$ denotes the number of out-of-sample forecasts, typical for macroeconomic applications. In both experiments we abstract from dynamics; independence is therefore a product of the design of the experiments induced by the random number generator. We also abstract from parameter uncertainty and assume sample estimates equal to their population counterparts; accordingly our results are insensitive to the size of any in-sample (training) period. This lets us focus on issues of specification rather than estimation.

5.5.1 Monte-Carlo Experiment #1

This experiment examines the accuracy of using KLIC weights, (26), to estimate the mixing coefficient w_i in (22). Knowing this true weight, we consider the accuracy of the KLIC based estimates as a function of various parameters in the DGP. Accuracy is measured by comparing the average (across R) KLIC estimate (and its variance) with the true weights.

Following (22) the true density is defined as the finite mixture:

$$f_t(y_t) = w_1 g_{1t}(y_t) + (1 - w_1) g_{2t}(y_t), \tag{28}$$

where $0 \leq w_1 \leq 1$ and $g_{1t}(y_t)$ and $g_{2t}(y_t)$ are two individual density forecasts:

$$g_{1t}(y_t) = N(\alpha x_{1t}, 1), \tag{29}$$
$$g_{2t}(y_t) = N(\beta x_{2t}, 1), \tag{30}$$

$(t = 1, ..., T)$, where x_{1t} and x_{2t} are uncorrelated randomly drawn N(1,1) variates. Then for each replication we compute $\{\widehat{w}_1\}_{r=1}^{R}$ using (26). For various values of w_1, α and β we then compute for $T = 50$: (i) $\overline{w}_1$, the average $\widehat{w}_1$ across R; (ii) σ_{w_1}, the standard deviation of $\widehat{w}_1$ across R and (iii) the Monte-Carlo standard error [MCSE] defined as $\sigma_{w_1}/\sqrt{R}$. σ_{w_1} measures the uncertainty of $\widehat{w}_1$ and MCSE measures the uncertainty of $\overline{w}_1$. When $g_{1t}(y_t) = g_{2t}(y_t) \Longrightarrow w_1 = 0.5 \Leftrightarrow w_1/w_2 = g_{1t}(y_t)/g_{2t}(y_t)$.

We draw two conclusions from the results of this Monte-Carlo experiment as summarised in Table 1. First, *ceteris paribus*, using (26) to estimate w_1 is most accurate for values of w_1 close to zero/unity. Secondly, *ceteris paribus*, the accuracy of the estimated weights increases the greater the distance between the two densities (the greater the value of $\beta - \alpha$). Accordingly, the most accurate estimates are delivered both when w_1 is close to zero/unity and when $\beta - \alpha \geq 2$, or in other words $\{y_t\}_{t=1}^{T}$ is generated exclusively by just one of the two densities, (29) or (30), and these two densities look very different. This is reflected by values of $\overline{w}_1$ close to w_1 and small values of σ_{w_1}. For mid-range values of w_1, even when $\beta - \alpha$ is large, there is considerable uncertainty about both $\widehat{w}_1$ and $\overline{w}_1$; while one cannot reject the hypothesis that $\overline{w}_1 = w_1$ there is considerable variation in $\widehat{w}_1$. This suggests that for small to moderate sized samples even when the data are generated

from a mixture density like (28), or (22), estimating the combination weights using (26) is unreliable unless one of the component densities *dominates* the mixture.

These two findings appear to reflect the limitations of using (4), and the maintained normality assumption, to estimate the KLICs, *via* (12); the estimated weights deviate most from the true weights when the combined density is most non-normal. This happens when the true weights are not close (or equal) to zero (or unity) and the distance between the two densities, as measured by $\beta - \alpha$, is large. In these two cases Berkowitz's LR test, as traditionally employed, cannot pick up the non-normality present under the DGP and accordingly delivers distorted weights. Future work should assess whether more accurate estimated weights are obtained using the modified Berkowitz test, suggested by Bao et al. (2004) and referred to in footnote 17, that relaxes the maintained normality assumption.

5.5.2 Monte-Carlo Experiment #2

This experiment (i) explores the size and power properties for the KLIC test of equal predictive performance (18) and (ii) examines and draws out some characteristics of density forecast combination using (22).[28] The latter lets us isolate some stylised situations in which KLIC based combination works, and some in which it does not.

As a comparator to the finite mixture (22) we consider as an alternative combined density forecast $N(m_t^*, v_t^*)$, the Gaussian density with mean equal to the combined point forecast, m_t^*, and variance v_t^* computed from the in-sample residuals (outturn less combined point forecast). Such a density forecast combination has been considered by Hendry & Clements (2004) and in a related meteorology literature by Déqué et al. (1994) and Wilks (2002).

Along the lines of Giacomini (2002) and Hendry & Clements (2004), our DGP is designed to reflect the case where two competing forecasters use misspecified models.[29] Each model is misspecified by excluding the variable the other model includes. As shown by Granger (1989) we know that while pooling the information is optimal, pooling the forecasts is not, in general. Specifically we consider the following DGP for the process $\{y_t\}$:

$$y_t = \alpha x_{1t} + \beta x_{2t} + \varepsilon_t, \tag{31}$$

where $\varepsilon_t \sim$ IID N(0,1), independently of the regressors, and the regressors x_{1t} and x_{2t} are uncorrelated N(0,1) variates. The true (conditional) density is therefore $f_t(y_t) = N(\alpha x_{1t} + \beta x_{2t}, 1)$. Two competing forecasters, unaware of the nature of the process governing the determination of y_t in (31), then compute density forecasts of y_t based on the following misspecified models:

$$y_t = \alpha x_{1t} + \varepsilon_{1t}, \tag{32}$$

$$y_t = \beta x_{2t} + \varepsilon_{2t}. \tag{33}$$

[28] Complimentary work by Giacomini (2002) has also investigated the size and power of the $KLIC$ test (as indicated in Section 4, under a scoring rules interpretation). We juxtapose this investigation with examination of the accuracy of the density forecasts individually as well as density forecast combination.

[29] As discussed by Hendry & Clements (2004) we remind ourselves that the reasons for success or failure of combination can be multi-faceted. Results using a specific DGP can only ever be illustrative.

Both correctly assume a normal density with a variance of unity. Their implied density forecasts are therefore, respectively:

$$\text{Forecast 1} \quad : \quad g_{1t}(y_t) = N(\alpha x_{1t}, 1), \tag{34}$$

$$\text{Forecast 2} \quad : \quad g_{2t}(y_t) = N(\beta x_{2t}, 1). \tag{35}$$

The pit's of each of these forecasts are:

$$z^*_{1t} = \varepsilon_t + \beta x_{2t}, \tag{36}$$

$$z^*_{2t} = \varepsilon_t + \alpha x_{1t}, \tag{37}$$

which are IID N(0,1) when $\beta = 0$ or $\alpha = 0$. The RMSE of the point forecast from (34) is $\sqrt{1+\beta^2}$ and that of forecast (35) is $\sqrt{1+\alpha^2}$. $E(z^*_{1t}) = 0$; $V(z^*_{1t}) = \sigma_1^2 = 1+\beta^2$. The expected difference in logarithmic scores $E(\ln g_{2t}(y_t) - \ln g_{1t}(y_t)) = (\beta^2 - \alpha^2)/2$ since $E(\ln(g_{2t}(y_t)) = -(\alpha^2+1)/2$ and $E(\ln(g_{1t}(y_t)) = -(\beta^2+1)/2$.[30] Therefore when $\alpha = \beta$ the two forecasts are equally accurate. As $\alpha = \beta$ rises both density forecasts are deteriorating, although equally so. We investigate the following as a function of α and β:

1. The rejection rates (across R) for $g_{1t}(y_t)$ and $g_{2t}(y_t)$ based on testing the statistical adequacy of the individual density forecasts $g_{1t}(y_t)$ and $g_{2t}(y_t)$ using the KLIC at 5%; see (11). Essentially these tabulate the size and power of the Berkowitz LR test (4) for the evaluation of an individual density forecast.

2. The rejection rate for the KLIC test of equal predictive performance; cf. (18). When $\alpha = \beta$ the the rejection rate is the size of the test. When $\alpha \neq \beta$ the rejection rate is the power of the test.

3. The average and variance (across R) of the KLIC for $g_{1t}(y_t)$ and $g_{2t}(y_t)$. Denote the averages for $g_{1t}(y_t)$ and $g_{2t}(y_t)$, KLIC_{m1} and KLIC_{m2}, respectively, with their variances denoted $\sigma^2_{\text{KLIC}_1}$ and $\sigma^2_{\text{KLIC}_2}$. The ratio of KLIC_1 to KLIC_2 indicates the relative gain in forecast accuracy delivered by using (33) rather than (32).

4. $\overline{w}_1$, the average and variance (across R) of the KLIC weights on $g_{1t}(y_t)$ in the BMA combined density forecast; see (26). The weight on $g_{2t}(y_t)$ is one minus the weight on $g_{1t}(y_t)$. Denote the average weight on $g_{1t}(y_t)$ by $\overline{w}_1$ and the variance by $\sigma^2_{w_1}$. We also consider the naïve BMA weights (27); these are denoted $\overline{w}_{1,naïve}$ and the variance $\sigma^2_{w_1,naïve}$.

5. The rejection rates for the BMA combined density forecast $p_t(y_t)$ using the KLIC weights based on testing the statistical adequacy of the combined density forecast using the KLIC at 5%. Results are also indicated using naïve weights.

6. The rejection rates for the density forecast $N(m^*_t, v^*_t)$ based on testing its statistical adequacy using the KLIC test at 5%.

[30] The R^2 based on the regression (32) is $1 - \frac{\beta^2+1}{\alpha^2+\beta^2+1}$; the R^2 based on the regression (33) is $1 - \frac{\alpha^2+1}{\alpha^2+\beta^2+1}$.

The results are summarised in Tables 2 and 3. Let us draw out five findings. First, *ceteris paribus*, the higher the ratio of KLIC_1 to KLIC_2, i.e. the more accurate Forecast 2 is relative to Forecast 1, the better the power of the KLIC test of equal predictive performance.

Secondly, power depends not just on this KLIC ratio but the accuracy of an individual density forecast. The less likely we are to reject an individual forecast (using LR1 or LR2), i.e. the more satisfactory an individual density forecast, the better the power for a given KLIC ratio. The test of equal predictive performance is also, in general, slightly over-sized although we found in unreported experiments that this tendency to over-reject vanishes for larger T. Use of the bootstrap test may also deliver improvement. These findings are consistent with Giacomini (2002).

Thirdly, the estimated weights $\overline{w}_1$ remain equal between the two competing density forecasts even when one model is much more accurate than the other, unless the best performing model is itself in absolute terms performing badly. This is explained, as discussed in Section 5, by the fact that for the weight to tend to zero the KLIC for the worst performing model (relative to the best model) must be very high. Therefore one can get equal weights for $[\alpha = 0; \beta = 1]$ despite the fact that model 2 is almost 6 times better as evidenced by the KLIC ratio just because the KLIC for Forecast 1 is not high enough. In our experiment one needs $\mathsf{KLIC}_{m1} - \mathsf{KLIC}_{m2} > 1$ for weights to budge from around 0.5, as seen for $[\alpha = 1, \beta = 2]$ and $[\alpha = 1, \beta = 3]$. Alternatively one can have a situation where both models are bad but as relatively speaking the better model performs well it is given a weight of one; e.g. see $[\alpha = 1, \beta = 3]$ where both models individually are rejected more than 88% of the time, but Forecast 1 receives a weight close to zero just because Forecast 2 is close to 20 times better as evidenced by their KLIC ratio.

Fourthly, the combined density forecast BMA, in general, delivers more accurate forecasts than the individual density forecasts. This is evidenced by a lower rejection rate for BMA than the individual density forecasts (rows LR1 and LR2). Combination using BMA performs quite well as long as α and β do not both get very high. Recall that the two-component mixture normal density with different means but common variances is symmetrical only if $w_1 = w_2 = 0.5$; see Everitt & Hand (1981), p.27. We are therefore probably overstating the validity of density forecast combination using BMA as the density evaluation test considered assumes normality, and will therefore not pick up any departures from it. Only when the weight equals zero [e.g. for $\alpha = 1, \beta = 3$] will the combined density forecast also be normal. In other cases the BMA density may become very non-normal while, by assumption, the true density is normal. Nevertheless, it is encouraging that the mean and variance of the transformed pit's z_{it}^* implied by the combined density forecast BMA have mean zero and a variance of unity more often than the individual density forecasts.

Fifthly, the evidence is mixed as to whether density forecast combination using naïve weights in BMA is better than using KLIC weights. As expected, the Gaussian combined density density forecast $N(m_t^*, v_t^*)$ does very well with rejection rates less than 1%.

6 Empirical Application: Comparing and Combining "Fan" Charts of UK Inflation

The application serves not just to illustrate density forecast comparison and combination but is also of considerable interest *per se*. Forecasting inflation is of pivotal importance for central banks in an era of inflation targeting. We focus on quarterly forecasts of one-year ahead RPIX inflation (RPI excluding mortgage payments: ONS code CHMK), the principal monetary policy target over the sample period. The year ahead forecasts correspond to a five quarter ahead horizon. Strictly the forecasts are conditional on the assumption that nominal interest rates remain constant throughout the forecast period; however, following previous analysis, we regard the forecasts as unconditional on the (plausible) assumption that inflation does not react within a year to changes in interest rates; see Clements (2004) and Wallis (2004).

6.1 The Bank of England "fan" chart

We consider the quarterly sequence of one-year ahead inflation forecasts published by the Bank of England. These forecasts are published in the *Inflation Report* in February, May, August and November, which we correspond to q1, q2, q3 and q4, respectively. The Bank of England has published density forecasts for RPIX inflation from 1993q1. Up until 1995q4 these took the form of charts showing the central projection, together with an estimate of uncertainty based on the historical mean absolute error. At this stage the Bank of England did not quantify a skew so that the mode, median and mean projections are equal; the density forecast is (implicitly) assumed normal.[31] From 1996q1 the Bank of England has published the so-called "fan" chart, that allows for skewness or the "balance of risks" to be on the upside or downside; see Britton et al. (1998). From 1997q3 these charts have been based on the deliberations of the Monetary Policy Committee (MPC).[32] The final projection for RPIX inflation, prior to the new target for inflation announced by the Chancellor in December 2003, was published in the February 2004 *Inflation Report*.

The fan chart is based analytically on the two-piece normal distribution; for details see Wallis (2004). The Bank of England publishes the parameter values underlying each published fan chart by supplying *via* its spreadsheets information on the following five statistics: the mode (μ^d), median, mean ($E(Y)$), uncertainty (σ) and skew. The uncertainty statistic is a parameter of the two-piece normal distribution; see Wallis (2004), Box A, for details - following Wallis note that we correct earlier confusion about what the uncertainty measure published by the Bank of England represents. The skew statistic is defined as the mean minus the mode. Given these parameters, following Wallis, we

[31] In 1995q1 uncertainty is not recorded; we simply assume the value from the previous *Inflation Report*. This seems a reasonable assumption given that uncertainty is being quantified based on historical RMSE which should not be expected, at least in large-samples, to change much quarter from quarter.

[32] The density forecasts from 1993q1-1997q2 are available at: http://www.bankofengland.co.uk/inflationreport/historicalforecastdata.xls. From 1997q3 they are available at http://www.bankofengland.co.uk/inflationreport/rpixinternet.xls.

can back-out the standard errors σ_1 and σ_2 of the two normal distributions on which the two-piece normal distribution is based. Then, see also Clements (2004), we can compute the pit as follows:

$$P(Y < y) = \left\{ \begin{array}{l} \frac{2\sigma_1}{\sigma_1+\sigma_2}\Phi\left(\frac{y-\mu^d}{\sigma_1}\right) \text{ for } y < \mu^d \\ \left(\frac{\sigma_1-\sigma_2}{\sigma_1+\sigma_2}\right) + \frac{2\sigma_2}{\sigma_1+\sigma_2}\Phi\left(\frac{y-\mu^d}{\sigma_2}\right) \text{ for } y > \mu^d \end{array} \right\}, \tag{38}$$

where Φ is the standard normal c.d.f..

6.2 The NIESR "fan" chart

We also consider the quarterly forecasts of annual RPIX inflation published in the *National Institute Economic Review*.[33] Since 1992q3 NIESR has, albeit in a sense implicitly, published probability forecasts for inflation, in that the *Review* contained the table "Average Absolute Errors". This table indicated the historical accuracy of NIESR forecasts by reporting the mean absolute error.[34] Since 1996q1 NIESR has explicitly published probability forecasts for inflation. These have taken the form of tabular histograms, indicating the probability of inflation falling within a band, although these bands have changed periodically. These probability forecasts are centred on the point forecast published in the *Review*. This point forecast is produced by NiGEM, a large-scale macroeconometric model. In deriving the density forecasts, normality is assumed. This is because earlier work that analysed the historical errors (from 1984-1995) made in forecasting RPI inflation could not reject normality; nor indeed could they reject unbiasedness (in fact rationality); see Poulizac et al. (1996). The variance of the density forecast is then set equal to the variance of the historical forecast error.[35]

The *Review* focuses on forecasting inflation in the fourth quarter of the current year and the fourth quarter of the next year; therefore only the q4 publication offers a one-year head forecast. While we can extract from back-issues of the *Review* one-year ahead point forecasts for the other quarters, published uncertainty estimates are only available for q4. Therefore, we follow Mitchell (2005) in his summary of National Institute density forecasts and make an assumption in order to infer uncertainty estimates for the other quarters. We simply assume the density forecast is normal with standard deviation equal across the four quarters in a year. This assumption is sensible if we believe NIESR only re-calibrated their forecast variances once a year.[36]

[33]The *Review* is currently published in January, April, July and October. Prior to 1996 the publication timetable was slighly different. In any case we refer to the four publications of the *Review* each year as q1, q2, q3 and q4. Given our interest in one-year ahead forecasts it does not seem unreasonable to ignore these changes to the publication timetable since the information set is little different and there is still one year's worth of shocks.

[34]Assuming normality, a 58% confidence interval around the point forecasts corresponds to the point estimate $\pm$ the mean absolute error.

[35]Past forecast errors are commonly used as a practical way of forecasting future errors; e.g. see Wallis (1989), pp. 55-56.

[36]The quarterly time-series of density forecasts used in this paper are available from Mitchell (2005).

Given the backward looking and mechanistic nature to NIESR's method of determining the variance, it is important what historical sample period is chosen to estimate the variance. Until 2002 NIESR considered historical forecast errors back until 1982. Since then they have focused on errors post 1993 and the variance of their density forecast dropped dramatically; see Mitchell (2005) for details. As we shall see below, with the advantage of hindsight we observe that by, until 2002, considering historical forecast errors back until 1982, NIESR were basing their uncertainty forecasts on their track-record across two different inflation 'regimes', the recent regime (post 1993/4) characterised by lower volatility. This serves as a timely reminder to forecasters that just as with point forecasts, basing density forecasts on past experience can lead to misleading forecasts when regimes change. NIESR was in fact well aware of this. For example, to quote from Poulizac et al. (1996) (p. 62), "Both our inflation forecast and the reliability of this forecast must depend on the seriousness with which the government approaches inflation targetting. It is not clear that past experience is a good guide to this... and, in turn, [this] probably implies that the error variances [based on historical performance]... overstate the current uncertainty associated with the inflation rate".[37] But until 2002 NIESR continued to publish uncertainty forecasts based on forecast errors back to 1982. However Mitchell (2005) has found that a break in the unconditional variance of NIESR's forecast errors around 1993-94 could have been detected *via* recursive analysis of these forecast errors towards the end of 1996.

6.3 Empirical results

In this section we present the results of evaluating, comparing and combining the Bank of England and NIESR point and density forecasts. Using actual inflation data up to 2004q2 we have 42 point and density forecasts in total to compare with the subsequent outturn for RPIX inflation from 1994q1-2004q2. We also present results based on forecasts made from 1997q3, as this corresponds to the period in which the MPC were charged with responsibility for the Bank of England's forecasts.

We consider the performance of, specifically, the combined density forecasts both in-sample and using recursive out-of-sample experiments. In-sample we compute the combining weights on the two forecasts using all of the 42 time-series observations. The out-of-sample analysis is designed to simulate whether in practice, in real-time, one could have pooled the Bank of England and NIESR density forecasts to obtain 'better' forecasts. Accordingly, from 1997q3 recursively we re-estimate the combination weights using data available up to period $(t-5)$. This acknowledges the fact that one has to wait five quarters to evaluate the performance of a given (year-ahead) forecast. These recursively computed optimal weights are then used to produce a series of combined density forecasts from

[37] NIESR has considered how stochastic simulation can be used as an alternative to historical errors to measure the uncertainty associated with the inflation rate; see Blake (1996). It is explained that this is expected to deliver a better measure of uncertainty if a new policy regime (say a new target for inflation) has been adopted. Using a coherent policy structure with interest rate setting determined by a monetary policy rule, Blake found that stochastic simulation suggested a smaller inflation standard error.

1997q4 to 2004q2.

As well as the actual density forecasts published by the Bank of England and NIESR, we consider a slight modification of each. For the Bank of England, we impose normality on the fan chart so that while the mean and variance of the fan chart are as published by the Bank of England, any skewness is ignored; we denote this forecast Bank_N. For NIESR we continue to consider a normal density forecast centred on the point forecast but rather than using the standard deviation estimates published by NIESR in q4 we compute rolling estimates. The standard deviation of the density forecast is computed recursively each quarter by setting it equal to the standard deviation of the pooled (across quarters) historical forecast errors (back to 1988) across a rolling 20 quarter window. At each point in time only information that would have been available in real-time is used; we denote this density forecast NIESR_R.[38] While not explicitly conditioning on any break in the variance, use of a rolling window is a simple means of ignoring some 'old' forecast error data.

The accuracy of the point forecasts is summarised by their RMSE. The accuracy of the density forecasts is summarised by the average log score, the KLIC and the p-value from the associated Berkowitz LR test LR_B which amounts to a test for whether $\mathsf{KLIC} = 0$. As indicated in Section 2, we expect autocorrelation in z^*_{it} since we are looking at five-step ahead quarterly forecasts; the forecast horizon is longer than the periodicity of data. Therefore, although for completeness we present KLIC and LR estimates based on $p_t(z^*_{it}) = \phi\left[\left(z^*_{it} - \mu - \rho z^*_{it-1}\right)/\sigma\right]/\sigma$, we focus on those that are unaffected by dependence in z^*_{it} and consider $p_t(z^*_{it}) = \phi\left[\left(z^*_{it} - \mu\right)/\sigma\right]/\sigma$; let KLIC_{IN} and LR_{IN} denote the former tests for zero mean, unit variance and zero first-order autocorrelation and KLIC_N and LR_N denote the latter tests for zero mean and unit variance, both under a maintained hypothesis of normality. Recall that $\mathsf{KLIC}_{IN} = LR_{IN}/2T$ and $\mathsf{KLIC}_N = LR_N/2T$, where LR_{IN} and LR_N are computed *via* substitution of $p_t(z^*_{it})$, defined appropriately, in (4) or (14).

6.3.1 Evaluation of point forecasts

Table 4 uses the RMSE to summarise the accuracy of the Bank of England and NIESR point forecasts. Over both the 1993q1-2003q2 and 1997q3-2003q2 periods the Bank of England's forecasts are more accurate than those of NIESR, as evidenced by a lower RMSE. However, as we see from Table 5, even over the full-sample period this difference is not statistically significant at 95% using a DM test with HAC estimation of the variance; the associated test statistic is 1.19.[39] Statistical tests (not reported) also reject the significance of the bias component of RMSE for both the Bank of England and NIESR, again using a HAC robust estimator of its standard error.

[38] The NIESR_R density forecasts used are available from Mitchell (2005).

[39] This result is robust to the small-sample correction suggested by Harvey et al. (1997).

6.3.2 Evaluation of "fan" charts

Table 4 also summarises the performance of the Bank of England and NIESR density forecasts. Examination of LR_{IN} and LR_N reveals that over the full-sample period the Bank of England density forecasts are rejected as statistically adequate (at best the p-value is 0.003). However, while one rejects the density forecasts using LR_{IN} over the 1997q3-2003q2 period, LR_N suggests the density forecasts are well specified (with a p-value of 0.097). This rejection using LR_{IN} should not be assumed to imply statistical inadequacy since this test may be contaminated by the serial correlation we expect in the pit's. LR_N, on the other hand, is robust to such dependence.

These results are consistent with earlier studies that tend to find that the Bank of England (year-ahead) density forecasts fail tests for independence (constituting no violation of forecast optimality) but perform better against the distributional ones, at least over the 1997q3- period.[40] Interestingly, if we impose normality on the Bank of England density forecasts, with mean and variance as before, accuracy is only marginally worse (indeed the KLIC estimates are identical to 2 d.p.); the Bank of England's assumption of a two-piece normal distribution since 1996q1, empirically at least, makes little practical difference.

In contrast Table 4 shows that the NIESR density forecasts, at least as actually published, are rejected by both LR_{IN} and LR_N over both sample periods. Again this is consistent with earlier work; see Hall & Mitchell (2004*a*). The distributional failure is not surprising since until 2002 NIESR clearly over-estimated the degree of uncertainty associated with its point forecast. As indicated, this is explained by their reliance on a mechanical examination of historical forecast errors too far back into the past. We note that the Bank of England and NIESR mean forecast errors have a correlation coefficient of 0.73.

However the re-calibrated NIESR density forecasts NIESR_R, that rely on rolling estimates of the density's standard deviation, do perform well. They pass the distributional test, LR_N, both over the full and shorter samples periods (with p-values of 0.44 and 0.53, respectively). Indeed the density forecast NIESR_R performs better than any of the Bank of England's density forecasts. Given the poorer performance of NIESR point estimates, this reminds us of the importance of using reliable uncertainty forecasts.

[40] Clements (2004) evaluates Bank density forecasts of year-ahead inflation, using a range of statistical tests, over the shorter sample period, 1997q3-2002q1. Hall & Mitchell (2004*a*) also provide a more detailed evaluation of Bank of England and NIESR density forecast, like Clements, considering a range of statistical tests. While we focus here on evaluation of density forecasts using the Berkowitz LR test, it should be noted, particularly given the small-sample size available, that exploratory data analysis based on examination of the plot against the uniform distribution and the auto-correlation functions does suggest the Bank (albeit, if our results are reliable, in a statistically insignificant manner) over-estimated the degree of uncertainty. The distribution function for the Bank is S-shaped. This indicates that they placed too much weight in the tails. This is consistent with the findings of Wallis (2004).

6.3.3 Comparing the Bank of England and NIESR "fan" charts

Table 5 presents the results of the KLIC tests that statistically compare the accuracy of the Bank of England and NIESR point and density forecasts. HAC robust estimators of the variance of the loss differential $\{d_t\}$ are used when undertaking these tests of equal forecast accuracy, specifically we use a Newey-West estimator with Bartlett weights.

As mentioned in Section 6.3.1 above, we cannot reject the null hypothesis of equal point forecast accuracy using a traditional DM test, assuming a quadratic loss function, with a test statistic of 1.19. However, using the test proposed in Section 4, we find that the Bank of England's fan chart is more accurate, in a statistically significant manner, than NIESR's (the test statistic is 10.01, with an associated asymptotic 95% critical value of 1.96). But NIESR's re-calibrated density forecasts NIESR_R are statistically better than the Bank's (the test statistic is 4.55).

Furthermore, although Table 4 indicated that imposing normality on the Bank of England density forecasts rendered them slightly less accurate, Table 5 reveals that one cannot reject the hypothesis that imposing normality makes no difference (the test statistic is 0.37). This implies that the Bank of England's opinion about the balance of risks is statistically no better than assuming there is no upside or downside risk to inflation.

6.3.4 Combining the Bank of England and NIESR "fan" charts

Tables 6 and 7 analyse the combined point and density forecasts. Table 6 presents the combination weights on Bank of England and NIESR, exploiting (in-sample) all of the available 42 time-series observations. Table 6 shows that the Bank of England receives a weight of 1.04 and NIESR a weight of -0.1 when optimally combining the two competing point estimates using a G-R type regression with no intercept. The robust standard errors also presented indicate that only the Bank of England's point forecast is statistically significant. Indeed statistically one can accept *encompassing*: the Bank of England has a weight of unity and NIESR a weight of zero. Similar results are obtained if one constrains the weights to sum to unity.

In contrast Table 6 shows that when combining the density forecasts, the weights on NIESR are more equal to those on the Bank of England. NIESR receives a weight of 0.43 while the re-calibrated NIESR density forecasts NIESR_R receives a weight of 0.53.

Table 7 then considers the accuracy of the combined point and density forecasts both in-sample and in the recursive out-of-sample experiments.

Let us consider the in-sample results for the point forecasts first. Optimal combination of point forecasts must reduce the RMSE. Reassuringly, Table 7 shows that this is indeed the case! The optimal combined point forecast has a RMSE of 0.497 as opposed to a RMSE of 0.53 for the Bank of England and 0.83 for NIESR (see Table 4). If we constrain the combination weights to sum to unity then the combined RMSE is still lower than the individual RMSEs but is slightly higher than the estimate presented in Table 7. Use of equal weights leads to less accurate point forecasts, as does use of BMA weights (these are the weights used to combine the density forecasts based on the KLIC estimates). Out-of-sample this result is reversed; equal and BMA weights perform similarly but beat optimal

weights (equal weights have a RMSE of 0.40 opposed to 0.46 for the recursively computed optimal weights). Explanations for this are put forward in Hendry & Clements (2004) and Smith & Wallis (2005).

A different picture emerges when we analyse the performance of the combined density forecasts. While in-sample combination of point forecasts (using optimal G-R weights) must deliver greater accuracy, in-sample combination of the density forecasts using BMA KLIC weights need not. Indeed this is what happens in Table 7. The combined density forecast using BMA weights is worse than the Bank of England density forecast (with a KLIC of 0.35 rather than 0.33).

Accordingly, Hall & Mitchell (2004*b*) propose that one combine density forecasts by performing a numerical search to choose that set of combination weights that deliver the best value of the test statistic used to evaluate the performance of density forecasts (in our case the Berkowitz LR test statistic). In other words, these weights minimise the 'distance', as measured by the KLIC, between the forecasted and true, but unknown, density. In this way, as with combination using the G-R regression, combination cannot deliver worse forecasts than the individual forecasts in-sample. It may, however, as happens in this application deliver a weight of unity on the Bank of England and zero on NIESR, essentially implying that combination cannot help.

Table 7 also shows that in-sample density forecast combination using BMA weights is better than use of equal weights. This is reflected by a higher score and lower KLIC estimates. However, which is particularly interesting in the light of Monte-Carlo experiment #2, density forecast combination using (22) is worse than just using a normal combined density forecast $N(m_t^*, v_t^*)$. Combination using $N(m_t^*, v_t^*)$ delivers satisfactory density forecasts (with a p-value of 0.65) that are also better than either the Bank of England or NIESR density forecast individually (see Table 4).

Out-of-sample the performance of the combined density forecast with BMA weights is still worse than the Bank of England's; this is reflected by a KLIC of 0.39 rather than the 0.1 for the Bank of England seen in Table 4. Use of BMA weights is, however, better than use of equal weights; the KLIC rises to 0.42. The normal combined density $N(m_t^*, v_t^*)$ again does better out-of-sample than density forecast combination using (22), with a KLIC of 0.13. But this is still less accurate than the Bank of England individually. Finally, we considered combining the Bank of England and NIESR$_R$ density forecasts. This does deliver density forecasts that pass the LR$_N$ test (with a p-value of 0.09). But one would have done better considering the NIESR$_R$ density forecast alone; again density forecast combination does not help.

6.3.5 Summing up the empirical results

Both the point and density forecasts published by the Bank of England are more accurate than those published by NIESR, and at least since 1997q3 appear to pass the evaluation test that KLIC $= 0$. The test for equal density forecast accuracy proposed in this paper also revealed the Bank of England's density forecast to be better statistically than NIESR's. However, the Bank of England's opinion about the balance of risks is statistically no

better than assuming there is no risk. Moreover, it was easy to construct, using real-time data, alternative variance estimates for NIESR's density forecast that result in NIESR's density forecast beating the Bank's.

We also found that the combined density forecasts did not beat the best individual density forecasts. This appears to be consistent with the view that combination of a 'good' forecast with an inferior forecast can make matters worse. However, the KLIC weights did deliver better combined density forecasts than equal weights. Clearly, as with the combination of point forecasts, the weights used matter.

7 Concluding Comments

This paper has proposed and analysed the Kullback-Leibler Information Criterion (KLIC) as a unified statistical tool to evaluate, compare and combine density forecasts. Computation of the KLIC is facilitated by exploiting its relationship with the well-known Berkowitz LR test for the evaluation of individual density forecasts based on the pit's.

We have found that the KLIC provides a useful (and statistically quite powerful) tool to compare competing density forecasts statistically. However, care should be exercised when using the finite mixture density to combine density forecasts. Although the KLIC provides a theoretically attractive means of weighting the competing densities in this mixture, since the best forecast according to the KLIC has the highest posterior probability, in contrast to point forecast combination with optimal weights density forecast combination using KLIC weights need not equal or improve upon the performance of the best individual forecast even in-sample yet alone out-of-sample. The finite mixture density can generate highly non-normal combined densities that may or may not help. We saw in the second Monte-Carlo experiment that combination can help but the empirical application revealed use of the finite mixture density made the density forecasts worse than the best individual density forecast. A Gaussian combined density forecast, centred on the combined point forecast, did better than the mixture normal combination. However, there is no reason to expect this will always be the case.

Table 1: Monte-Carlo Experiment #1

w_1	α	β	$\overline{w}_1$	σ_{w_1}	MCSE
0	0.00	0.00	0.50	0.03	0.001
0	0.00	1.00	0.31	0.07	0.003
0	0.00	2.00	0.04	0.03	0.001
0	0.00	3.00	0.00	0.00	0.000
0	1.00	1.00	0.32	0.07	0.003
0	1.00	2.00	0.04	0.03	0.001
0	1.00	3.00	0.00	0.00	0.000
0.2	0.00	0.00	0.50	0.03	0.002
0.2	0.00	1.00	0.39	0.07	0.003
0.2	0.00	2.00	0.14	0.10	0.004
0.2	0.00	3.00	0.03	0.07	0.003
0.2	1.00	1.00	0.38	0.07	0.003
0.2	1.00	2.00	0.13	0.10	0.004
0.2	1.00	3.00	0.03	0.06	0.003
0.4	0.00	0.00	0.50	0.03	0.001
0.4	0.00	1.00	0.46	0.08	0.003
0.4	0.00	2.00	0.36	0.19	0.008
0.4	0.00	3.00	0.27	0.26	0.012
0.4	1.00	1.00	0.46	0.08	0.004
0.4	1.00	2.00	0.34	0.18	0.008
0.4	1.00	3.00	0.28	0.28	0.013
0.6	0.00	0.00	0.50	0.03	0.001
0.6	0.00	1.00	0.54	0.08	0.004
0.6	0.00	2.00	0.66	0.19	0.009
0.6	0.00	3.00	0.74	0.27	0.012
0.6	1.00	1.00	0.54	0.08	0.003
0.6	1.00	2.00	0.63	0.18	0.008
0.6	1.00	3.00	0.73	0.27	0.012
0.8	0.00	0.00	0.50	0.03	0.001
0.8	0.00	1.00	0.62	0.07	0.003
0.8	0.00	2.00	0.87	0.10	0.004
0.8	0.00	3.00	0.97	0.06	0.003
0.8	1.00	1.00	0.62	0.08	0.003
0.8	1.00	2.00	0.86	0.10	0.004
0.8	1.00	3.00	0.97	0.08	0.004
1	0.00	0.00	0.50	0.03	0.001
1	0.00	1.00	0.69	0.06	0.003
1	0.00	2.00	0.96	0.03	0.001
1	0.00	3.00	1.00	0.00	0.000
1	1.00	1.00	0.69	0.07	0.003
1	1.00	2.00	0.96	0.03	0.001
1	1.00	3.00	1.00	0.00	0.000

Notes: w_1 is the true weight in (28); α and β control the characteristics of the two component densities (29) or (30); $\overline{w}_1$ is the mean of $\{\widehat{w}_1\}_{r=1}^{R}$; σ_{w_1} is the standard deviation of $\{\widehat{w}_1\}_{r=1}^{R}$ and $MCSE$ is the Monte-Carlo standard error, where $MCSE = \sigma_{w_1}/\sqrt{R}$.

Table 2: Monte-Carlo experiment #2

α	0.000	0.000	0.000	0.000	0.000	0.100	0.100	0.100	0.100
β	0.000	0.200	0.400	0.700	1.000	0.100	0.580	1.015	1.65
LR_1	0.084	0.048	0.086	0.382	0.876	0.052	0.200	0.858	1.00
LR_2	0.084	0.050	0.042	0.060	0.064	0.044	0.052	0.052	0.050
$KLIC_{m1}$	0.033	0.030	0.037	0.076	0.187	0.029	0.052	0.183	0.764
σ_{KLIC_1}	0.027	0.024	0.029	0.055	0.102	0.025	0.041	0.101	0.277
$KLIC_{m2}$	0.033	0.030	0.030	0.031	0.032	0.029	0.031	0.030	0.031
σ_{KLIC_2}	0.027	0.024	0.023	0.025	0.027	0.024	0.025	0.026	0.025
KLIC equal	0.000	0.114	0.240	0.620	0.910	0.052	0.472	0.904	1.00
$\overline{w}_1$	0.500	0.500	0.498	0.489	0.462	0.500	0.495	0.462	0.327
σ_{w_1}	0.000	0.003	0.006	0.013	0.025	0.002	0.010	0.026	0.059
$\overline{w}_{1,naive}$	0.500	0.501	0.450	0.315	0.166	0.501	0.395	0.163	0.044
$\sigma_{w_1,naive}$	0.000	0.112	0.177	0.202	0.141	0.069	0.217	0.146	0.042
BMA	0.084	0.046	0.040	0.060	0.060	0.046	0.048	0.044	0.076
BMA naive	0.084	0.044	0.036	0.044	0.046	0.046	0.038	0.046	0.078
$N(m_t^*, v_t^*)$	0.010	0.004	0.006	0.010	0.010	0.002	0.002	0.006	0.006

Notes: LR_1 and LR_2 denote the rejection rates, at 5%, for testing the optimality of the individual density forecasts $g_{1t}(y_t)$ and $g_{2t}(y_t)$ using the Berkowitz LR test (4); $KLIC_{m1}$ and $KLIC_{m2}$ are the average values of the KLIC, (12), for $g_{1t}(y_t)$ and $g_{2t}(y_t)$ across Monte-Carlo replications R, and σ_{KLIC_1} and σ_{KLIC_2} are the standard deviations across replications; KLIC equal is the rejection rate, at 5%, for the KLIC test of density forecast equality (18); $\overline{w}_1$ and σ_{w_1} are the average and standard deviation (across R) of the KLIC weights on $g_{1t}(y_t)$ in the BMA combined density forecast, cf. (26); $\overline{w}_{1,naive}$ and $\sigma_{w_{1,naive}}$ are the average and standard deviation (across R) of the KLIC weights on $g_{1t}(y_t)$ in the naive BMA combined density forecast, cf. (27); BMA is the rejection rate, at 5%, for testing the optimality of the combined density forecast $p_t(y_t)$ with KLIC weights (26) using the Berkowitz LR test (4); similarly BMA naive and $N(m_t^*, v_t^*)$ are the rejection rates for the combined density forecast using, respectively, naive BMA weights (27) and the Gaussian combined density forecast with mean equal to the optimal combined point estimate and variance calculated from the in-sample residuals using m_t^*.

Table 3: Monte-Carlo experiment #2 (cont.)

α	0.200	0.200	0.200	0.200	0.400	0.400	1.000	1.000	1.000	1.000
β	0.200	0.400	0.700	1.000	0.400	1.000	1.000	1.400	2.000	3.000
LR_1	0.046	0.098	0.374	0.874	0.100	0.846	0.842	0.988	1.000	1.000
LR_2	0.050	0.060	0.058	0.062	0.100	0.106	0.822	0.892	0.884	0.886
$KLIC_{m1}$	0.029	0.037	0.074	0.193	0.036	0.178	0.181	0.466	1.232	3.343
σ_{KLIC_1}	0.023	0.031	0.054	0.107	0.029	0.101	0.104	0.203	0.413	0.923
$KLIC_{m2}$	0.029	0.033	0.030	0.032	0.036	0.036	0.183	0.189	0.185	0.186
σ_{KLIC_2}	0.025	0.027	0.025	0.027	0.029	0.031	0.107	0.105	0.103	0.107
KLIC equal	0.068	0.138	0.540	0.862	0.068	0.698	0.060	0.304	0.908	1.000
$\overline{w}_1$	0.500	0.499	0.489	0.460	0.500	0.465	0.500	0.432	0.268	0.056
σ_{w_1}	0.003	0.006	0.013	0.025	0.006	0.025	0.032	0.051	0.075	0.045
$\overline{w}_{1,naive}$	0.489	0.475	0.318	0.160	0.496	0.184	0.501	0.296	0.136	0.056
$\sigma_{w_1,naive}$	0.128	0.172	0.207	0.133	0.195	0.144	0.176	0.132	0.073	0.034
BMA	0.046	0.058	0.048	0.060	0.050	0.064	0.090	0.114	0.188	0.552
BMA naive	0.044	0.044	0.038	0.054	0.046	0.062	0.084	0.140	0.234	0.502
$N(m_t^*, v_t^*)$	0.002	0.004	0.000	0.004	0.002	0.006	0.008	0.002	0.006	0.006

Table 4: Individual Forecast Performance: Point and Density Results

	1993q1-2003q2						1997q3-2003q2					
	RMSE	$\overline{S}$	KLIC_{IN}	KLIC_N	LR_{IN}^p	LR_N^p	RMSE	$\overline{S}$	KLIC_{IN}	KLIC_N	LR_{IN}^p	LR_N^p
Bank	0.53	-0.83	0.33	0.14	0.000	0.003	0.40	-0.61	0.25	0.10	0.007	0.097
Bank_N	0.53	-0.83	0.33	0.14	0.000	0.003	0.40	-0.62	0.25	0.10	0.008	0.095
NIESR	0.83	-1.57	0.91	0.43	0.000	0.000	0.45	-1.42	0.86	0.75	0.000	0.000
NIESR_R	0.83	-1.21	0.20	0.02	0.001	0.435	0.45	-0.90	0.12	0.03	0.127	0.534

Notes: $\overline{S}$=Average Log Score; KLIC_{IN} is the KLIC estimate for zero mean, unit variance and zero first-order dependence derived by scaling LR_{IN} by 2T; KLIC_N is the KLIC estimate for zero mean and unit variance derived by scaling LR_N by 2T; Bank_N is Bank of England fan chart with normality imposed; NIESR_R is NIESR density forecast based on rolling estimates of the standard deviation; LR_{IN}^p is the p-value for the three degrees of freedom (Berkowitz) LR test for zero mean, unit variance and zero first-order dependence LR_{IN}; LR_N^p is the p-value for the two degrees of freedom (Berkowitz) LR test for zero mean and unit variance LR_N

Table 5: Comparing Bank of England and NIESR point and density forecasts 1993q1-2003q2

DM: Bank=NIESR	1.189
KLIC equal: Bank=NIESR	10.007
KLIC equal: Bank=NIESR$_R$	4.551
KLIC equal: Bank=Bank$_N$	0.374

Notes: DM is a Diebold-Mariano test for equality between the Bank of England and NIESR point forecasts. KLIC equal is the test for equal density forecast accuracy between the Bank of England and NIESR density forecasts, and modified versions of each; see (18). The asymptotic 95 per cent critical value is 1.96.

Table 6: Combination weights on the Bank of England and NIESR point and density forecasts: 1993q1-2003q2

		w_1	w_2
Point		1.042	-0.106
	robust e.s.e.	(0.213)	(0.227)
Density		0.572	0.428
Density: Bank vs. NIESR$_R$		0.470	0.530

Notes: The point forecasts are combined using a G-R regression; e.s.e. is estimated standard error, in parentheses; w_1 denotes the weight on the Bank of England density and w_2 the weight on the NIESR density.

Table 7: Combined Forecast Performance: Point and Density Results

	In-sample (1993q1-2003q2)			Out-of-sample (1997q3-)		
Point Forecasts	RMSE			RMSE		
Optimal weights	0.497			0.457		
Equal weights	0.642			0.397		
BMA weights	0.619			0.393		
Density Forecasts	KLIC$_N$	$\overline{S}$	LR$_N^p$	KLIC$_N$	$\overline{S}$	LR$_N^p$
BMA weights	0.345	-1.076	0.000	0.392	-0.894	0.000
$N(m_t^*, v_t^*)$	0.010	-0.719	0.654	0.131	-0.676	0.043
BMA: Equal weights	0.373	-1.125	0.000	0.422	-0.894	0.000
BMA weights: Bank and NIESR$_R$	0.133	-0.995	0.004	0.090	-0.720	0.114

References

Bao, Y., Lee, T.-H. & Saltoglu, B. (2004), A test for density forecast comparison with applications to risk management. Department of Economics, UC Riverside.

Bates, J. M. & Granger, C. W. J. (1969), 'The combination of forecasts', *Operational Research Quarterly* **20**, 451–468.

Berkowitz, J. (2001), 'Testing density forecasts, with applications to risk management', *Journal of Business and Economic Statistics* **19**, 465–474.

Blake, A. (1996), 'Forecast error bounds by stochastic simulation', *National Institute Economic Review* **156**, 72–79.

Bomberger, W. (1996), 'Disagreement as a measure of uncertainty', *Journal of Money, Credit and Banking* **28**, 381–392.

Britton, E., Fisher, P. & Whitley, J. (1998), 'The inflation report projections: understanding the fan chart', *Bank of England Quarterly Bulletin* **38**, 30–37.

Burnham, K. P. & Anderson, D. R. (2002), *Model selection and multimodel inference: A practical information-theoretic approach. (Second edition)*, Springer-Verlag: New York.

Clemen, R. & Winkler, R. (1999), 'Combining probability distributions from experts in risk analysis', *Risk Analysis* **19**, 187–203.

Clements, M. P. (2003), 'Editorial: Some possible directions for future research', *International Journal of Forecasting* **19**, 1–3.

Clements, M. P. (2004), 'Evaluating the Bank of England density forecasts of inflation', *Economic Journal* **114**, 844–866.

Clements, M. P. (2005), 'Evaluating the Survey of Professional Forecasters probability distributions of expected inflation based on derived event probability forecasts', *Empirical Economics* . Forthcoming.

Clements, M. P. & Hendry, D. F. (1998), *Forecasting Economic Time Series*, Cambridge University Press: Cambridge.

Clements, M. P. & Smith, J. (2000), 'Evaluating the forecast densities of linear and non-linear models: Applications to output growth and unemployment', *Journal of Forecasting* **19**, 255–276.

Corradi, V. & Swanson, N. R. (2004), 'Predictive density and conditional confidence interval accuracy tests', *Journal of Econometrics* . Forthcoming.

Déqué, M., Royer, J. F. & Stroe, R. (1994), 'Formulation of gaussian probability forecasts based on model extended-range integrations', *Tellus* **46A**, 52–65.

Diebold, F. X., Gunther, A. & Tay, K. (1998), 'Evaluating density forecasts with application to financial risk management', *International Economic Review* **39**, 863–883.

Diebold, F. X. & Mariano, R. S. (1995), 'Comparing predictive accuracy', *Journal of Business and Economic Statistics* **13**, 253–263.

Diebold, F. X., Tay, A. S. & Wallis, K. F. (1999), Evaluating density forecasts of inflation: the Survey of Professional Forecasters, *in* R. Engle & H. White, eds, 'Cointegration, causality and forecasting: a festschrift in honour of Clive W. J. Granger', Oxford University Press.

Draper, D. (1995), 'Assessment and propagation of model uncertainty', *Journal of the Royal Statistical Society, Series B* **57**, 45–97.

Everitt, B. S. & Hand, D. J. (1981), *Finite Mixture Distributions*, London: Chapman and Hall.

Fernandez-Villaverde, J. & Rubio-Ramirez, J. (2004), 'Comparing dynamic equilibrium economies to data: A Bayesian approach', *Journal of Econometrics* **123**, 153–187.

Garratt, A., Lee, K., Pesaran, M. H. & Shin, Y. (2003), 'Forecast uncertainties in macroeconometric modelling: an application to the UK economy', *Journal of the American Statistical Association* **98**, 829–838.

Genest, C. & Zidek, J. (1986), 'Combining probability distributions: a critique and an annotated bibliography', *Statistical Science* **1**, 114–135.

Giacomini, R. (2002), Comparing density forecasts via weighted likelihood ratio tests: asymptotic and bootstrap methods. UCSD Discussion Paper 2002-12.

Giacomini, R. & White, H. (2004), Tests of conditional predictive ability. Department of Economics, University of California, San Diego.

Giordani, P. & Söderlind, P. (2003), 'Inflation forecast uncertainty', *European Economic Review* **47**, 1037–1059.

Good, I. J. (1952), 'Rational decisions', *Journal of the Royal Statistical Society, Series B* **14**, 107–114.

Granger, C. W. J. (1989), 'Combining forecasts - twenty years later', *Journal of Forecasting* **8**, 167–173.

Granger, C. W. J. & Jeon, Y. (2004), 'Thick modeling', *Economic Modelling* **21**, 323–343.

Granger, C. W. J. & Pesaran, M. H. (2000), 'Economic and statistical measures of forecast accuracy', *Journal of Forecasting* **19**, 537–560.

Granger, C. W. J. & Ramanathan, R. (1984), 'Improved methods of combining forecasts', *Journal of Forecasting* **3**, 197–204.

Granger, C. W. J., White, H. & Kamstra, M. (1989), 'Interval forecasting: an analysis based upon ARCH-quantile estimators', *Journal of Econometrics* **40**, 87–96.

Hall, S. G. & Mitchell, J. (2004*a*), Density forecast combination. National Institute of Economic and Social Research Discussion Paper No. 249.

Hall, S. G. & Mitchell, J. (2004*b*), "Optimal" combination of density forecasts. National Institute of Economic and Social Research Discussion Paper No. 248.

Hamilton, J. D. (1994), *Time Series Analysis*, Princeton University Press.

Harvey, D. L., Leybourne, S. J. & Newbold, P. (1997), 'Testing the equality of prediction mean square errors', *International Journal of Forecasting* **13**, 273–281.

Hendry, D. F. & Clements, M. P. (2004), 'Pooling of forecasts', *Econometrics Journal* **7**, 1–31.

Li, F. & Tkacz, G. (2001), A consistent bootstrap test for conditional density functions with time-dependent data. Bank of Canada Working Paper No. 2001-21.

Mitchell, J. (2005), 'The National Institute density forecasts of inflation', *National Institute Economic Review* **193**, 60–69.

Morris, P. (1974), 'Decision analysis expert use', *Management Science* **20**, 1233–1241.

Morris, P. (1977), 'Combining expert judgments: A Bayesian approach', *Management Science* **23**, 679–693.

Pesaran, M. H. & Zaffaroni, P. (2004), Model averaging and value-at-risk based evaluation of large multi asset volatility models for risk management. University of Cambridge.

Poulizac, D., Weale, M. & Young, G. (1996), 'The performance of National Institute economic forecasts', *National Institute Economic Review* **156**, 55–62.

Raftery, A. E., Madigan, D. & Hoeting, J. A. (1997), 'Bayesian model averaging for linear regression models', *Journal of the American Statistical Association* **92**, 179–191.

Sarno, L. & Valente, G. (2004), 'Comparing the accuracy of density forecasts from competing models', *Journal of Forecasting* **23**, 541–557.

Smith, J. & Wallis, K. F. (2005), Combining point forecasts: the simple average rules, OK? Department of Economics, University of Warwick.

Stock, J. & Watson, M. (2004), 'Combination forecasts of output growth in a seven-country data set', *Journal of Forecasting* **23**, 405–430.

Tay, A. S. & Wallis, K. F. (2000), 'Density forecasting: a survey', *Journal of Forecasting* **19**, 235–254.

Vuong, Q. (1989), 'Likelihood ratio tests for model selection and non-nested hypotheses', *Econometrica* **57**, 257–306.

Wallis, K. F. (1989), 'Macroeconomic forecasting: a survey', *Economic Journal* **99**, 28–61.

Wallis, K. F. (2004), 'An assessment of Bank of England and National Institute inflation forecast uncertainties', *National Institute Economic Review* **189**, 64–71.

Wallis, K. F. (2005), 'Combining density and interval forecasts: a modest proposal', *Oxford Bulletin of Economics and Statistics* . This Issue ?

West, K. D. (1996), 'Asymptotic inference about predictive ability', *Econometrica* **64**, 1067–1084.

White, H. (1984), *Asymptotic theory for econometricians*, Academic Press: Orlando, Florida.

White, H. (2000), 'A reality check for data snooping', *Econometrica* **68**, 1097–1126.

Wilks, D. S. (2002), 'Smoothing forecast ensembles with fitted probability distributions', *Quarterly Journal of the Royal Meteorological Society* **128**, 2821–2836.

Winkler, R. (1981), 'Combining probability distributions from dependent information sources', *Management Science* **27**, 479–488.

Zarnowitz, V. & Lambros, L. (1987), 'Consensus and uncertainty in economic prediction', *Journal of Political Economy* **95**, 591–621.